DOCUMENT RESUME

ED 283 875 TM 870 414

AUTHOR	Cook, Linda L.; And Others
TITLE	Characteristics of Samples and Linking Items Affecting a Partial Pre-Calibrations Design.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	College Entrance Examination Board, New York, N.Y.
PUB DATE	Apr 87
NOTE	51p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*College Entrance Examinations; Concurrent Validity; *Equated Scores; Estimation (Mathematics); Higher Education; *Item Analysis; Item Sampling; *Latent Trait Theory; Mathematical Models; Mathematics Tests; Statistical Studies; Test Reliability; *Transformations (Mathematics); Verbal Tests
IDENTIFIERS	*Calibration; LOGIST Computer Program; *Scholastic Aptitude Test

ABSTRACT

This study tests several explanations for discrepant results in an earlier study (Cook et al., 1985) which presented a partial pre-calibration method for equating new editions of the Scholastic Aptitude Test (SAT) to the same scale as older editions. In contrast to full pre-calibration, which seeks to equate all items from two or more editions, the partial method designates a subset of items to serve as an equating section; both methods rely on item response theory (IRT) and employ LOGIST to achieve calibration. Several verbal and mathematics equatings selected from the larger study were subjected to further tests. Estimation error of item parameters was small; instead, the source of difficulty rested with the IRT characteristic curve transformations, which failed to adequately equate the series of calibrations runs to the same scale. Possible differences in ability levels at different test administrations could not account for the discrepant equatings. Finally, the study identified items that functioned differently (DIF items) in the two groups used for equating and calibration. Eliminating these items from the partial pre-calibrations runs, however, did not improve the equatings significantly. Figures depict the calibration designs, item response functions, and other results. (LPG)

Characteristics of Samples and Linking Items Affecting a Partial Pre-calibration Design[1,2,3]

Linda L. Cook
Daniel R. Eignor
Marilyn S. Wingersky

Educational Testing Service

[1]A paper presented at the annual meeting of AERA, Washington, 1987.

[2]This project was supported by the College Board through Admissions Testing Program funding.

[3]The authors would like to acknowledge the assistance or advice of Nancy Petersen, Fred McHale, Nancy Wright, Karen Carroll, and Ted Blew in completing this project.

Characteristics of Samples and Linking Items Affecting a Partial Pre-calibration Design

Linda L. Cook
Daniel R. Eignor
Marilyn S. Wingersky

Educational Testing Service

INTRODUCTION

Since 1982, IRT equating of new editions of the Scholastic Aptitude Test (SAT) has been based, except for a small number of instances, on three-parameter logistic model item parameter estimates (see Lord, 1980) obtained from the concurrent calibration of items from the new edition, two equating tests, and two old editions of the test, using data from two samples taking the new edition of the test and a sample from each group taking the old editions of the test (see Figure 1). In a concurrent calibration design, item parameters for the three total tests and the two equating tests are estimated and placed on a common scale in a single calibration run. The computer program LOGIST (Wingersky, Barton, and Lord, 1982; Wingersky, 1983) has been used to perform the item calibration needed. Scores on the new edition are then equated to scores on each of the earlier editions, using IRT true-score equating (Lord, 1980) and the results averaged. This type of IRT equating uses exactly the same data collection design that was used for the traditional non-IRT equating of SAT-verbal and SAT-mathematical done prior to 1982. The calibration design is based on the SAT braiding plan (Angoff, 1974) and is considerably limited in its flexibility. Scores on the new test edition can only be equated to scores on old editions that were administered with the same equating sections as those given with the new edition. A more flexible equating procedure would take advantage of item

parameter estimates from test editions given at a number of different SAT administrations that are on one common scale.

Insert Figure 1 about here

The most flexible calibration design that could be used with the SAT would be a full pre-calibration design, which would lead to pre-equating of the verbal and mathematical sections. Pre-equating refers to the process of establishing a conversion from raw to scaled scores prior to the time the new test edition is administered operationally. The process depends on the adequate pretesting of a pool of items from which the new test edition will be assembled, the calibration of these items using IRT methods, and the utilization of a linking scheme to place the IRT item parameter estimates on a common scale. The last step is, perhaps, the most critical step. Unlike the concurrent calibration design, where the necessary item parameter estimates are automatically on the same scale because there is only one calibration run, for the pre-calibration design, there will be multiple calibration (LOGIST) runs and the parameter estimates will initially be on the unique scales defined by the ability distributions of the samples used in the separate LOGIST runs (see Cook and Eignor, 1983). It is possible, however, if the IRT model fits the data, and there are common items between calibration runs, to determine a linear relationship that can be used to transform item parameter estimates from one calibration run to the scale of the parameter estimates from another calibration run. Hence, it is not the existence of unique scales that presents a problematic hurdle for IRT pre-equating but, rather, the need to have sets of common items between calibration runs that would ultimately allow placement of <u>all</u> pretest parameter estimates for items constituting final editions of a test on one common scale. Further, feasibility

studies investigating the possibility of pre-equating the SAT (Eignor, 1985; Eignor and Stocking, 1986) have provided results that have been, for the most part, unacceptable. For these reasons, a somewhat less flexible calibration design than pre-equating, but certainly more flexible than the concurrent calibration design presently used, was seen as worthy of investigation. This design, which is called partial pre-calibration, is described first in the following paragraphs, and then the results of a feasibility study (Cook, McHale, Eignor, Petersen, and Dorans, 1985) investigating the possibility of its use are described. The current investigation involves further study of selected results from the Cook et al. (1985) study.

The essential feature of a partial pre-calibration design is that the items from the equating test have been calibrated and placed on a common scale prior to their administration with a new edition of the SAT. (In full pre-calibration, all the items in the new edition have been calibrated prior to the administration and placed on a common scale, not just an equating section.) In performing the equating, data is collected from the sample who take the new edition and also the equating test, for which IRT parameters have been previously estimated. The parameter estimates for the equating items, which are recalibrated with the new edition and which already exist on the common scale from a previous calibration, provide the link necessary to place new edition item parameter estimates on the common scale. With the existence of multiple equating sections containing items on the common scale comes a degree of flexibility not offered by the concurrent calibration design. A distinct advantage of the partial pre-calibration design is that equating sections are interchangeable; any equating section with items on the common scale can be administered with the new edition for equating purposes, not just those equating sections that were given with the old editions to be used

in the equating as is the case for the concurrent calibration design based on the SAT braiding plan. (Note that in order for partial pre-calibration to work any old editions that might be used in the equating must also contain items on the common scale.) In addition to this greater degree of flexibility, there is a considerable cost savings associated with the equating of the new edition in that the old editions to be used in the equating do not have to be recalibrated with the new edition, as is presently the practice.

Once the new edition parameter estimates are placed on the common scale, multiple equatings to all old editions with item parameters on the common scale become possible. In the concurrent design, which uses equating plans laid out in the SAT braiding plan, equating to only the two old editions that were administered with the common item material is possible. The use of multiple old editions in the equating process should ultimately improve upon current equating practice and make scores more consistent from one administration to another.

As mentioned previously, Cook et al. (1985) conducted a feasibility study investigating the possibility of using a partial pre-calibration design to equate new editions of the SAT. In their study, item parameters needed for the equatings were either estimated through a number of individual LOGIST calibration runs done specifically for the study or obtained from previous concurrent calibration runs performed in the context of operational IRT equating. Each of the calibrations, be they concurrent calibrations previously done or calibrations done specifically for the study, produced item parameter estimates on a scale particular to the calibration run. Parameter estimates from the separate calibrations were then placed on one common scale. Among the equating tests calibrated in any given calibration run in this study was an equating test that was calibrated in another run. Thus, two sets of parameter estimates existed for

these items that were on different scales because they resulted from two different calibration runs. The characteristic curve transformation procedure (Stocking and Lord, 1983) was used to place the item parameter estimates from the "current" run on the scale of the parameter estimates obtained in the previous run. For the study, an extensive design was developed that permitted placing item parameter estimates for 24 SAT-verbal editions and 24 SAT-mathematical editions on a common scale (one for SAT-verbal, one for SAT-mathematical) defined in November 1982 when a particular edition of the SAT, designated E8, was first administered. Items from equating sections that were calibrated along with the different editions of the SAT were also placed on this scale. Further details and pictorial representations of these calibration designs may be found in the Appendix of this paper.

Once the calibrations were completed and all item parameter estimates were placed on the E8 base scale, it was possible to equate the scores from any particular edition of the SAT to the scores from any other edition of the SAT used in the study. For purposes of the study, the test to be equated was treated as if it had never been equated previously. IRT true-score equating (Lord, 1980) was then carried out to the same two old editions that were used when the new edition was equated operationally through a concurrent calibration design. Hence, an appropriate criterion existed in all cases against which to compare the results of the experimental equatings; i.e., the operational equatings resulting from the concurrent calibrations which were used for score reporting. As mentioned previously, the study design also permitted equating each new edition to multiple (more than two) old editions. However, the maximum number of old editions used for the equatings in the Cook et al. (1985) study was two.

The results obtained from the equatings in the Cook et al. study were somewhat difficult to interpret. Some of the equatings agreed very closely with the criterion equatings based on the concurrent calibration design and some produced quite discrepant results. Cook et al. also found it difficult to conclude, based on the residual plots and tabulations they prepared, whether or not a partial pre-calibration design appeared to be more feasible for SAT-verbal than for SAT-mathematical. For both tests, there were a number of equatings that produced residuals greater than 20 scaled-score points. These results led Cook et al. to question very seriously the implementation of a partial pre-calibration design for either SAT-verbal or SAT-mathematical.

A search for possible explanations for why some equatings produced smaller residuals than others in the Cook et al. study was not particularly fruitful. For instance, whether or not new or old editions of a test were linked to the base scale by a single transformation or by several transformations, prior to equating, seemed to have little effect. Efforts to evaluate the effect of particular equating tests that were used more than once also resulted in conflicting information. Cook et al. concluded their paper by listing a number of additional factors that could have possibly affected their partial pre-calibration results; several of these factors are investigated in detail in this paper. In summary, Cook et al. felt that it did not seem unreasonable to hope that if some of the factors affecting the viability of the partial pre-calibration design could be determined and controlled, equatings based on partial pre-calibration would eventually provide reasonable results.

PURPOSE

The purpose of this study was to investigate, for selected equatings from the Cook et al. (1985) study, factors that could have possibly affected their

partial pre-calibration results. First, a series of item parameter transformations and equatings were carried out in an attempt to determine if the poor results from the previous study were related to the item parameter transformations or to the item calibrations. Secondly, two factors that could influence the transformation process were examined: 1) possible differences in the ability levels of the samples of examinees used to calibrate the items used for linking purposes, and 2) the possible existence of differential item functioning of the linking items for the samples used to calibrate the data. The results of two recently completed studies (Stocking and Eignor, 1986; Cook, Eignor, and Wingersky, 1987) suggest that these two factors are viable candidates for explaining the poor partial pre-calibration results in the Cook et al. (1985) study.

Stocking and Eignor (1986) provided simulation results that have implications for the process of equating when there are large differences in ability between equating samples. These researchers were interested in what effects the ability levels of the samples used would have on three-parameter logistic model parameter estimates and subsequent equating results. Using LOGIST for estimation purposes, they found that differences in mean true ability between samples used in equating can cause differences in the precision with which parameters are estimated, even when the test data fit the particular model used. The effect of this differential precision in estimation on test equating can be substantial if the samples begin to have fairly large differences in true ability (i.e., a difference in means of one or more standard deviations on the ability scale). A more detailed explanation for why differences in mean true abilities can cause differences in the precision of the parameter estimates can be found in Stocking and Eignor (1986; pp. 11-13).

Cook, Eignor, and Wingersky (1987) investigated the effects on equating results of common item sections that contained a few items whose item response functions were not well fit by the three-parameter logistic model and of common item sections that contained a few items on which the two groups taking the common item section responded differently. Their study was also carried out using the three-parameter logistic item response theory model and Monte Carlo procedures. So that the simulated data reflected actual test data, the true item parameters were taken from the estimated parameters obtained from LOGIST calibrations of item responses obtained from selected administrations of the verbal sections of the SAT. These effects were investigated using both the concurrent calibration and the characteristic curve transformation linking procedures for varying numbers of common items. The effects of these common item sections were studied using a uniform ability distribution and certain characteristics of the parameter estimates for the common items (i.e., the parameter estimates for the items all had small standard errors of estimation) selected as a result of the findings of a previous study (Wingersky, Cook, and Eignor, 1986).

Results of the Cook et al. (1987) study indicated that equatings, particularly those obtained using a characteristic curve transformation design, are seriously affected by the presence of linking items that function differently for the two groups used to provide data for the equating and calibration. This is particularly true for shorter linking tests. They concluded that the quality of an equating depends, somewhat, on prior screening of linking tests and removal of items that function differently for the two groups.

METHODOLOGY

Choice of Equatings to Study

From the verbal and mathematical experimental equatings (i.e., equatings based on a partial pre-calibration design) in the Cook et al. (1985) study, one verbal and one mathematical equating was chosen for further study. Each of these equatings actually involved a pair of single equatings that were averaged, i.e., a pair is an equating of one new test edition to two old editions. The specific verbal and math equatings were chosen because: 1) one equating in the pair that were averaged gave excellent results when compared to the operational concurrent calibration criterion equating while the second equating in the pair gave extremely discrepant results, and 2) the parameter estimates for the new edition to be equated in each case came from the operational concurrent calibration run from which equating results used for score reporting were derived. This allowed some special analyses, described in detail later in the paper, to be developed. These analyses were used to explore the calibration runs in an effort to determine if the discrepant equatings resulted from problems with the estimation process or with the item transformation process.

In Figure 2, the equating relationships among the editions chosen for further study are depicted; these relationships are defined in the SAT braiding plan (Angoff, 1974). Upper case letters and numbers designate operational editions; lower case letters designate equating sections. The equating sections depicted in Figure 2 are those used in the concurrent calibration of the new and old editions.

Insert Figure 2 about here

Figure 3 contains portions of the SAT-verbal and SAT mathematical partial pre-calibration transformation plans presented in Cook et al. (1985) and in the Appendix of this paper. In each case, the portion depicted contains the specific editions being investigated in the current study. In the Cook et al. (1985) study, for both SAT-verbal and mathematical, all parameter estimates for the editions to be equated were transformed to the scale defined by Edition E8 (run 1 in Figure 3), using the characteristic curve transformation method (Stocking and Lord, 1983), and then the equatings were performed. For SAT-verbal, the E7 to C5 equating, after placing all parameter estimates on the E8 scale, gave inferior results when compared to the criterion equating from the concurrent calibration while for SAT-mathematical, the F3 to E3 equating gave inferior results. Figure 4 contains residual or difference plots for the four individual verbal and mathematical equatings being studied. In each case, raw (formula) score differences (partial pre-calibration results minus concurrent calibration criterion results) are shown for the range of possible raw scores.

Insert Figures 3 and 4 about here

Exploration of Calibration Runs

In an attempt to explore possible explanations for the differences in quality of the individual SAT-verbal and SAT-mathematical partial pre-calibration equatings under study, additional item transformations and equatings, making use of data from the Cook et al. (1985) study, were performed. Most of the equatings and transformations made use of parameter estimates for the editions involved that had already been placed on the base scale (run 1 in Figure 3) as part of the previous study (Cook et al., 1985).

Item Parameter Estimate Transformations

The first set of experimental transformations and equatings carried out for this study were performed in an attempt to assess how well the transformations used for the partial pre-calibration study (Cook et al., 1985), placed the item parameter estimates for editions used in the equatings on the same scale. Reference to Figure 3 will be useful in understanding the description that follows. In the previous study, SAT-verbal old edition B7 was calibrated in two separate LOGIST runs, run 2 (which also contained new edition E7) and run 4; parameter estimates obtained in both runs were then placed on the scale defined by run 1. The same sort of situation existed for old edition C5 in that it was calibrated separately in runs 2 and 3. After parameter estimates for test editions calibrated in runs 2, 3 and 4 were placed on the scale of run 1, new edition E7 (calibrated in run 2) was equated to old editions C5 and B7 (calibrated in runs 3 and 4, respectively).

One was to assess whether or not parameter estimates for the separate calibration runs shown in Figure 3 were adequately placed on their respective verbal or mathematical run 1 scales is to compare the transformed parameter estimates for, say, verbal edition C5 as it appears in runs 2 and 3. That is, compare the transformed parameter estimates for this edition resulting from the specific transformations carried out for the partial pre-calibration study. To make comparisons such as the one just described, a series of additional item parameter transformations and test equatings were carried out.

First, referring to the SAT-verbal part of Figure 3, the characteristic curve transformation procedure (Stocking and Lord, 1983), the same procedure that was used for the partial pre-calibration study to place runs 2-4 on the scale of run 1, was used to place runs 3 and 4 directly on the scale of run 2. For this

experimental transformation, all 85 items in edition B7 (or C5) were used as the linking test. Next, the linear parameters derived from the transformation of B7 item parameter estimates in run 4 to the scale of B7 parameter estimates in run 2 were examined, along with the linear parameters obtained from the transformation of item parameter estimates for edition C5 in run 3 to the scale of item parameter estimates for edition C5 appearing in run 2. If the item parameter estimates of the respective test editions were on scale together as a result of the partial pre-calibration study transformations, the linear parameters of the transformations obtained by this "direct link" approach should be very close to those of a 45° line, i.e., a line with a slope of one and an intercept of zero.

Similar transformations were carried out to investigate the partial pre-calibration results for the selected SAT-mathematical editions described in Figure 3. All special transformations carried out for this portion of the study are summarized in Table 1.

Insert Table 1 about here

In addition to the transformations described above, special equatings were carried out to gather additional information regarding whether or not the transformations used for the partial pre-calibration study succeeded in placing item parameter estimates for the separate calibration runs on their respective verbal or mathematical run 1 scales. Referring again to the verbal portion of the calibration scheme depicted in Figure 3, test editions B7 and C5 were equated to themselves using one set of item parameter estimates that were the result of the previously described direct link transformations and a second set of item parameter estimates that were the result of the transformations carried out for the partial pre-calibration study. For example, edition B7 appearing in run 4

that had been placed directly on the scale of run 2 (parameter estimates resulting from the direct link transformation; subsequently referred to as run 4*) was equated to edition B7 appearing in run 4 (parameter estimates resulting from the transformations carried out from the partial pre-calibration study). Similar equatings to those carried out for the verbal editions were carried out for mathematical editions C3 and E3. These equatings are referred to in Table 1 as special equatings 1.

The results of special equatings 1 were interpreted in the following manner. If the transformations carried out for the partial pre-calibration study resulted in placing the calibration runs on their respective verbal or mathematical run 1 scales, special equatings 1 should provide equating relationships represented by a 45° line. Any deviation from the expected results of equating a test to itself were interpreted as indicating a problem with the item parameter transformations developed in the partial pre-calibration study.

The final set of equatings that were carried out to examine the transformations resulting from the partial pre-calibration study are referred to in Table 1 as direct link equatings. These equatings did not involve equating an edition to itself, but rather involved equating a new edition to an old edition of the test (e.g., verbal edition E7 to edition C5). For the new editions of the test, the equatings used run 2 item parameter estimates that were a result of the transformations carried out for the partial pre-calibration study. Transformed parameter estimates for the old test editions used in these direct link equatings were obtained by the transformations described previously, which used the entire 85 item verbal or 60 item math test to place item parameter estimates directly on the respective run 2 scales. The direct link equating results were then compared to the equating results derived using the concurrent calibrations

and partial pre-calibration study transformations of the same new and old editions.

The following example should clarify how the direct link equatings were carried out and how the results were interpreted. Consider SAT-mathematical new edition F3 (run 2) and old edition E3 (run 5), where the E3 parameter estimates have additionally been placed on the scale of run 2 by the direct link transformation, i.e., E3 (run 5)*. An equating of F3 to E3 under these conditions (referred to as a direct link equating) can be compared to the partial pre-calibration equating of F3 to E3 done in the Cook et al. (1985) study and to the Cook et al. criterion concurrent calibration equating of F3 to E3. If the partial pre-calibration results are the outlier, this can be taken as a further indication that the linkage of E3 in run 5 to the base scale (run 1) was inadequate. Equatings such as the one just described for the mathematical new edition F3 and old edition E3 were also carried out for mathematical new edition F3 and old edition C3 and for verbal new edition E7, equated to old editions C5 and B7, respectively.

Errors of Estimation

Another possible source of the discrepant partial pre-calibration study equating results, obtained for the equating of verbal edition E7 to edition C5 and the equating of mathematical edition F3 to edition E3, is errors of estimation for the item parameters calibrated in the separate LOGIST runs. The equatings designated in Table 1 as special equatings 2 were carried out in an attempt to explore the possibility of estimation errors.

Referring again to the verbal portion of the calibration scheme depicted in Figure 3, test editions B7 and C5 were equated to themselves using one set of item parameter estimates that were the result of the previously described direct

link transformations and a second set of parameter estimates resulting from the transformations carried out for the partial pre-calibration study. For example, edition C5 appearing in run 3 that has been placed directly on the scale of run 2 (parameter estimates resulting from the direct link transformations and subsequently referred to as run 3*) was equated to edition C5 appearing in run 2 (parameter estimates resulting from partial pre-calibration study transformations). These equatings should result, once again, in a 45° line, if the direct link transformation is viable and if errors of estimation did not seriously affect the parameter estimates of the items in the test edition as it appears in the separate calibration runs. Since the direct link transformations are based on a single transformation using a linking test containing 85 items, it seems reasonable to assume that any discrepancy from a 45° line obtained by the equatings is related to estimation errors in the two calibration runs of interest. Special equatings 2 were carried out for the two verbal and two mathematical equatings investigated in this study.

Differences in Ability Levels of Samples

As mentioned previously, Stocking and Eignor (1986) demonstrated the effect that the ability levels of samples used in three-parameter logistic model calibrations can have on subsequent IRT equating results. The results of the Stocking and Eignor (1986) study may be of relevance in explaining the Cook et al. (1985) poor partial pre-calibration results. If, for example, the ability levels of the groups used in calibration runs 2-4, for the verbal test editions identified in Figure 3, are widely disparate from the ability level of the group in calibration run 1, then these differences may be large enough to cause problems for the calibration procedures used. For SAT-mathematical, an additional relevant comparison would involve comparing the ability level of the

group in calibration run 5 to that of run 4. Raw score means and standard deviations on the common item linking sections between the calibration runs identified in Figure 3 were used to provide an indication of possible differences in ability levels of the samples.

Differential Item Functioning in Linking Tests

As mentioned earlier, Cook et al. (1987) demonstrated the effect on IRT equatings of contamination of linking item sets through the presence of a few linking items that functioned differently (DIF items) for the two groups used to provide data for equating and calibration. In the Cook et al. (1985) study, the presence of DIF items in the common item linking sections could have affected equating results for the partial pre-calibration equatings as well as for the criterion concurrent calibration equatings. If DIF items were present and did have an effect on partial pre-calibration results, one would suspect more such items, or items exhibiting extreme differences, for the SAT-verbal and mathematical partial pre-calibration equatings that provided inferior results, i.e., E7 to C5 for verbal and F3 to E3 for mathematical.

For the partial pre-calibration runs, two separate sets of parameter estimates exist for each linking item from each of the separate calibrations. Plots of the item characteristic curves (with parameter estimates on a common scale) from the separate calibrations were obtained. In addition, as a measure of the discrepancy between the item characteristic curves estimated in the separate calibrations for each common item, a weighted mean absolute difference (MAD) value was calculated. Using all individuals in the larger of the two samples taking each linking item, the absolute difference between the two item response functions for each person (i.e., value of $\hat{\theta}$) was obtained and then averaged over individuals.

Referring to Figure 3, the following linking sections were studied using the above described plot and index. For SAT-verbal, the common item sections linking runs 2-4 to run 1 (gw linking 2 to 1, gs linking 3 to 1, and gw linking 4 to 1) were studied. For SAT-mathematical, the common item sections linking runs 2-4 to run 1 (gh linking 2 to 1, gh linking 3 to 1, hf + gt linking 4 to 1) and run 5 to run 4 (gj) were studied. It should be noted that hf + gt constitutes a pooled linking section containing twice the number of items (50) than is contained in the usual SAT-mathematical equating or common item section. Reasons for using a pooled linking section to link these runs can be found in Cook et al. (1985).

RESULTS

Exploration of Transformation Runs

Item Parameter Estimate Transformations

Table 2 contains the linear parameters obtained from the previously described direct link transformations. The two verbal transformations consisted of placing item parameter estimates obtained in calibration runs 3 and 4 on the scale of run 2 (see Figure 3) using the 85 items contained in either edition C5 or edition B7 as the linking test. Similarly, the transformations carried out for the two SAT-mathematical editions consisted of placing the parameter estimates obtained in calibration runs 3 and 5 on the scale of run 2 using the 60 item E3 and C3 editions as linking tests.

Insert Table 2 about here

The information provided in Table 2 indicates that verbal edition B7 and mathematical edition C3, appearing in runs 4 and 3 respectively, were very nearly on the same scale as these same editions appearing in the verbal and mathematical

calibration runs 2; i.e., the slopes and intercepts of the linear transformations are close to one and zero. On the other hand, linear parameters obtained for the transformation runs that placed verbal edition C5 and mathematical edition E3 directly on the scales of the verbal and math run 2 calibrations indicate that these editions were not adequately placed on their respective run 1 scales by the partial pre-calibration study transformations. This information leads one to the conclusion that the transformations carried out for the partial pre-calibration study, designed to place the item parameter estimates for verbal calibration run 3 and mathematical calibration run 5 on the scale of their respective run 1 calibrations, were not successful.

Figure 5 contains difference plots for the two types of special equatings that involved equating a test edition to itself; these were described earlier in the text and in Table 1. Only the results of special equatings 1 are relevant for the present discussion. Special equatings 1 involve equating an old edition of SAT-verbal or SAT-mathematical to itself using item parameter estimates that are the result of the transformations carried out for the partial pre-calibration study and parameter estimates that are a result of the direct link transformations. The difference plots contain discrepancies (in raw score units) between special equating results and the identity transformation (special equating results minus identity transformation) for the full range of possible raw scores.

Insert Figure 5 about here

The difference plots contained in Figure 5 for special equatings 1 are designed to assess how well the partial pre-calibration study transformations placed item parameter estimates for the editions used in the respective equatings

on the same scale. These plots show very different results for the old editions involved in problematic partial pre-calibration equatings (C5 for SAT-verbal, E3 for SAT-mathematical) than for the old editions involved in the non-problematic partial pre-calibration equatings (B7 for SAT-verbal, C3 for SAT-mathematical). As can be seen from examination of these plots, equating editions C5 and E3 to themselves resulted in fairly large residuals when compared to the identity transformation. In contrast, residuals obtained from equating editions B7 and C3 to themselves were quite small. The plots provide a clear indication (as did the previously described transformation runs) that the editions used in the problematic partial pre-calibration equatings were not adequately placed on their respective run 1 scales by the transformations that were carried out for that study.

Figure 6 contains difference plots for the final set of equatings summarized in Table 1, referred to as direct link equatings. The four equatings shown in Figure 6 represent SAT-verbal new edition E7 equated to old editions C5 and B7 and SAT-mathematical new edition F3 equated to old editions C3 and E3. Recall, parameter estimates for the direct link equatings were placed on scale by a single transformation using the respective 85 item verbal or 60 item mathematical test edition as the linking test. The direct link equatings are compared to equatings obtained using parameter estimates placed on scale by transformations carried out for the partial pre-calibration study and also to the criterion concurrent calibration equatings.

Insert Figure 6 about here

Examination of the difference plots shown in Figure 6 reveals that the equatings based on the direct link transformations agree very closely with the

criterion concurrent calibration equatings. The outlier equatings are clearly the SAT-verbal E7 to C5 equating and the SAT-mathematical F3 to E3 equating based on the partial pre-calibration study transformations. These results not only confirm previous evidence that the transformations for the partial pre-calibration study failed to place verbal editions E7 and C5 and mathematical editions F3 and E3 on scale together, they also (by their close agreement with the equatings based on the concurrent calibrations) substantiate the use of the concurrent calibration equatings as criterion equatings for the study.

Errors of Estimation

As mentioned previously, one possible source of the discrepant results obtained by equating verbal edition E7 to C5 and mathematical edition F3 to edition E3 might be errors of estimation that occurred during item calibration. To explore this possibility, special equatings 2 (see Table 1) were carried out. Recall, special equatings 2 involved equating a test edition to itself using one set of item parameter estimates that were the result of the direct link transformations and a second set of parameter estimates resulting from the transformations carried out for the partial pre-calibration study. These equatings should result in a 45° line, if the direct link transformations are viable and if errors of estimation did not seriously effect the parameter estimates of the items in the test edition as it appears in the separate calibration runs.

Figure 5 contains difference plots resulting from equating a test to itself for each of the old editions used in this study; verbal editions C5 and B7 and mathematical editions C3 and E3. The results of interest for this discussion are those showing a comparison of special equatings 2 to the zero residual line. As can be seen from an inspection of the plots, the residuals resulting from special

equatings 2 are very small, indicating close agreement between the item parameter estimates obtained for the respective test editions as they appeared in the separate calibration runs. The results of special equatings 2 can be interpreted as indicating that estimation error is not a plausible explanation for the poor equatings obtained for editions E7 to C5 and F3 to E3 in the partial pre-calibration study.

Ability Levels of Samples

One possible explanation for the poor results obtained for the transformations carried out for the partial pre-calibration study might be differences in the ability levels of samples used to calibrate linking items administered with the new and old test editions. Table 3 presents summary data on performance on equating sections used to provide the links between adjacent calibration runs in the sections of the SAT-verbal and SAT-mathematical partial pre-calibration linkage systems being studied. Means and standard deviations are reasonably similar on the equating sections for groups used in the separate calibrations, with one notable exception. Mean performance on SAT-verbal equating section gw is quite different for the groups involved in calibration runs 1 and 4--about a third of a standard deviation different. In the Stocking and Eignor (1986) study, at around this level of difference in ability the researchers began to note some small differences in equating results due to differences in the precision with which the parameters were estimated. Hence, the Stocking and Eignor results could prove useful in explaining the poor Cook et al. (1985) partial pre-calibration results except for the fact that equating section gw, connecting calibration runs 1 and 4, provides the link that places old edition B7 on the base scale (run 1). As seen in Figure 4, the partial pre-calibration equating of new SAT-verbal edition E7 to old edition B7 provided

excellent results when compared to the concurrent criterion. For the linking sections involved in the inadequate partial pre-calibrations (gs linking runs 1 and 3 for SAT-verbal, hf + gt linking runs 1 and 4 and gj linking runs 4 and 5 for SAT-mathematical), ability levels of the two groups used in the calibrations and subsequent linkings were reasonably similar. In sum, it would appear that differences in the ability levels of the samples used in calibration and linking is not a major contributing factor to the poor partial pre-calibration equating results under study.

Insert Table 3 about here

Differential Item Functioning in Linking Tests

Since differential item functioning (DIF) was found to be a major factor in the adequacy of transformations carried out for the study by Cook, Eignor and Wingersky (1987), the presence of DIF in the common item linking sections was studied for the partial pre-calibration equatings. For this aspect of the study, major emphasis was placed on the weighted mean absolute difference (MAD) index in deciding on which items exhibited DIF to a degree that removal from the linking section seemed reasonable.

Figures 7 and 8 contain ordered stem and leaf diagrams of mean absolute differences (MAD) between item response functions for the SAT-verbal and SAT-mathematical equating sections after application of the partial pre-calibration study transformations. Careful consideration of the distributions of these MAD values, in conjunction with plots of the item characteristic curves derived from the two calibrations for each linking item, led to the decision that a MAD value greater than .035 would provide an

indication of DIF. This cut-off is represented by the dotted line in the stem and leaf diagrams in Figures 7 and 8.

Insert Figures 7 and 8 about here

Use of the .035 cut-off as an indication of DIF led to the identification of a small number of SAT-verbal and SAT-mathematical items that should possibly be removed from the partial pre-calibration linking sections. The item response functions for these items (on the same scale), derived from the separate calibrations, are presented in Figures 9 and 10. Consistent with expectations, sections linking editions exhibiting poor partial pre-calibration equating results (gs linking runs 1 and 3 for SAT-verbal, hf+gt linking runs 1 and 4 for SAT-mathematical) contained a greater number of DIF items than did sections linking editions that provided acceptable partial pre-calibration equating results.

Insert Figures 9 and 10 about here

The study of the presence of DIF items in the concurrent calibrations proved to be a difficult task in that only item parameter estimates based on the combined group of examinees responding to the linking items were available (see Figure 1). Thus, a summary index, such as MAD, could not be calculated. Because no procedure that paralleled the one employed to remove DIF items from the linking tests used for the partial pre-calibration transformation runs could be applied to the concurrent calibration linking sections, a decision was made not to rerun any of the criterion concurrent calibrations with items removed. The

implications of this decision will be discussed further in the conclusion section.

Partial Pre-calibration Equatings with DIF Items Removed

Items in the linking sections connecting partial pre-calibration equating calibration runs having MAD values greater than the .035 cut-off were removed from the linking sections and the characteristic curve transformation runs were redone. Figure 11 presents difference plots for the partial pre-calibration equatings with DIF items removed (referred to as current partial pre-calibration equatings) along with the previous partial pre-calibration results and the direct link equating results. The criterion in these plots is again the concurrent calibration equating results.

Insert Figure 11 about here

For the poor partial pre-calibration equatings (E7 to C5 for SAT-verbal, F3 to E3 for SAT-mathematical), removal of DIF items resulted in modest reductions in raw score differences when compared to the criterion concurrent calibration equating results. In other words, the current partial pre-calibration equating results provide only a slight improvement over previous results that were considered problematic. For the acceptable partial pre-calibration equatings from the Cook et al. (1985) study (E7 to B7 for SAT-verbal, F3 to C3 for SAT-mathematical), removal of DIF items did not improve the partial pre-calibration results much at all and in some places on the score scale, differences from the criterion concurrent calibration equatings were increased slightly.

CONCLUSION

The purpose of this study was to investigate factors that may have led to poor partial pre-calibration equating results in a previous study (Cook et al., 1985) and to attempt to improve on the poor partial pre-calibration results. First, a series of item parameter transformations and equatings were carried out in an attempt to determine if the poor results from the previous study were related to the item parameter transformations or calibrations. Convinced that the problem equatings were a result of the transformations, the authors then focused on 1) possible differences in the ability levels of the samples used to calibrate the items used for linking purposes, and 2) the possible existence of differential item functioning of the linking items for the samples used to calibrate the data.

Examination of summary performance data for samples taking sections linking editions exhibiting poor partial pre-calibration results led to the conclusion that ability differences were not a major contributor to poor equating results. Removal of DIF items from sections linking editions exhibiting poor partial pre-calibration equating results led to only modest improvements in these results when compared to the concurrent calibration criterion equatings. These results seem to run counter to those observed by Cook, Eignor, and Wingersky (1987) when these researchers simulated DIF items in common item linking sections. However, the DIF items used in the Cook, Eignor, and Wingersky study exhibited much greater differences in item parameter estimates (and, hence, much larger MAD values) than the items assumed to be demonstrating DIF in this study. In sum, either or both of these factors do not appear to be the sole contributors to the poor partial pre-calibration equating results in the Cook et al. (1985) study.

One criticism of this study lies in the fact that no items potentially exhibiting DIF were removed from the concurrent calibrations and subsequent equating results. Had such items been located and removed, differences between the partial pre-calibration equating results (with DIF items removed) and concurrent calibration equating results may have decreased more. However, it should be recalled that the direct link equatings, based on entirely different linking tests, agreed closely with the concurrent calibration equatings. Thus, it seems reasonable to assume that new and old editions were placed on scale properly by the concurrent calibrations (in spite of any DIF items that might have been present) and that equating discrepancies from this criterion indicate poor results because the partial pre-calibration transformations carried out with and without removal of DIF items did not result in the editions being on scale together.

The results of this study, which indicate that removal of DIF items from a linking or equating test do not substantially improve equating results, are difficult to accept. It has long been common practice to inspect items for DIF and to remove those exhibiting substantial differences when carrying out conventional equating procedures. Cook and Petersen (in press) summarize research conducted to investigate properties of linking items and how these properties affect equating results. They conclude, using research by a variety of investigators, that one must be very careful that linking tests represent, as much as possible, identical tasks for groups of examinees who take the new and old editions of a test to be equated, i.e., the presence of differential item functioning can have a serious effect on equating results.

A question of interest for the present study then is, why didn't removal of the items exhibiting DIF have a substantial effect on the equating results?

There are a number of possible explanations that can be offered for the results. First, perhaps the procedures used to detect DIF were inadequate. Recall, DIF was determined by the use of the MAD index and corroborated by visual inspection of plots of item response functions. It is possible that a non-IRT approach, such as the use of the Mantel-Haenszel (Holland and Thayer, 1986) or standardization (Dorans and Kulick, 1986) statistics, may provide a better means of identifying DIF items.

Secondly, the study carried out by Cook, Eignor and Wingersky (1987) examined the effect of including two items each exhibiting a substantial amount of DIF in a single direction (i.e., both items biased against the same group) in the linking test. Perhaps a more likely occurrence would be a small number of items exhibiting DIF, but in opposite directions, thus having a neutralizing effect on each other. Or, a typical situation may be a fairly large number of items, each exhibiting small but consistent DIF in a single direction, and hence having a cumulative effect on the transformation and, ultimately, the equating results. It is possible that either of these two situations existed for linking tests used for the partial pre-calibration study, but the procedures decided upon detect DIF were not adequate to isolate such occurrences.

To summarize, the results of the specific investigations carried out in the present study did not provide an explanation for the poor equating results obtained in the partial pre-calibration study. This is particularly disappointing since the results of two simulation studies (Cook, Eignor and Wingersky, 1987; Eignor and Stocking, 1986) strongly indicated that the problematic results could be related to either differences in ability levels of samples used to calibrate the linking tests or the presence of DIF items in the linking tests or both.

The results of the study do provide an indication that the inadequate partial pre-calibration equating results from the Cook et al. (1985) study were related to the fact that editions used in the problematic equatings were not on scale together as a result of application of the characteristic curve transformation procedure. It is apparent that some other factor or factors besides ability level differences and the presence of DIF items must be influencing these transformations. Given that the characteristic curve transformation procedure is a frequently used procedure by IRT practitioners involved in equating and differential item functioning studies, this is clearly an area where further research should be emphasized.

/kad
DE\AERACHAR

References

Angoff, W. H. (1974). The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests. New York: College Entrance Examination Board.

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, B. C.: Educational Research Institute of British Columbia.

Cook, L. L., & Petersen, N. S. (in press). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. Applied Psychological Measurement.

Cook, L. L., Eignor, D. R., & Wingersky, M. S. (1987, April). The effect on IRT equating of using linking items with problematic item response functions. Paper presented at the annual meeting of AERA, Washington.

Cook, L. L., McHale, F. J., Eignor, D. R., Petersen, N. S., & Dorans, N. J. (1985, April). Item response theory equating of aptitude tests: A partial pre-calibration design. Paper presented at the annual meeting of AERA, Chicago.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23, 355-368.

Eignor, D. R. (1985). An investigation of the feasibility and practical outcomes of pre-equating the SAT verbal and mathematical sections (RR-85-10). Princeton, NJ: Educational Testing Service.

Eignor, D. R., & Stocking, M. L. (1986). An investigation of possible causes for the inadequacy of IRT pre-equating (RR-86-14). Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. (1986). Differential items functioning and the Mantel-Haenszel procedure (RR-86-31). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Stocking, M. L., & Eignor, D. R. (1986). The impact of different ability distributions on IRT pre-equating (RR-86-49). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, B.C.: Educational Research Institute of British Columbia.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST V user's guide. Princeton, NJ: Educational Testing Service.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1986, April). Specifying the characteristics of linking items used for the item response theory item calibration. Paper presented at the annual meeting of AERA, San Francisco.

	New Edition	Equating Test A	Equating Test B	Old Edition A	Old Edition B
Sample 1	X	X			
Sample 2	X		X		
Sample 3		X		X	
Sample 4			X		X

Figure 1: Concurrent calibration designs for SAT-verbal and SAT-mathematical. (The "Xs" indicate the tests taken by the respective samples.)

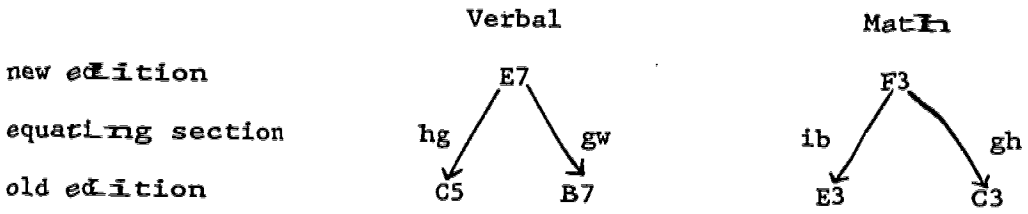


Figure 2: Depiction of equating relationships among the SAT-verbal and SAT-mathematical editions chosen for further study.

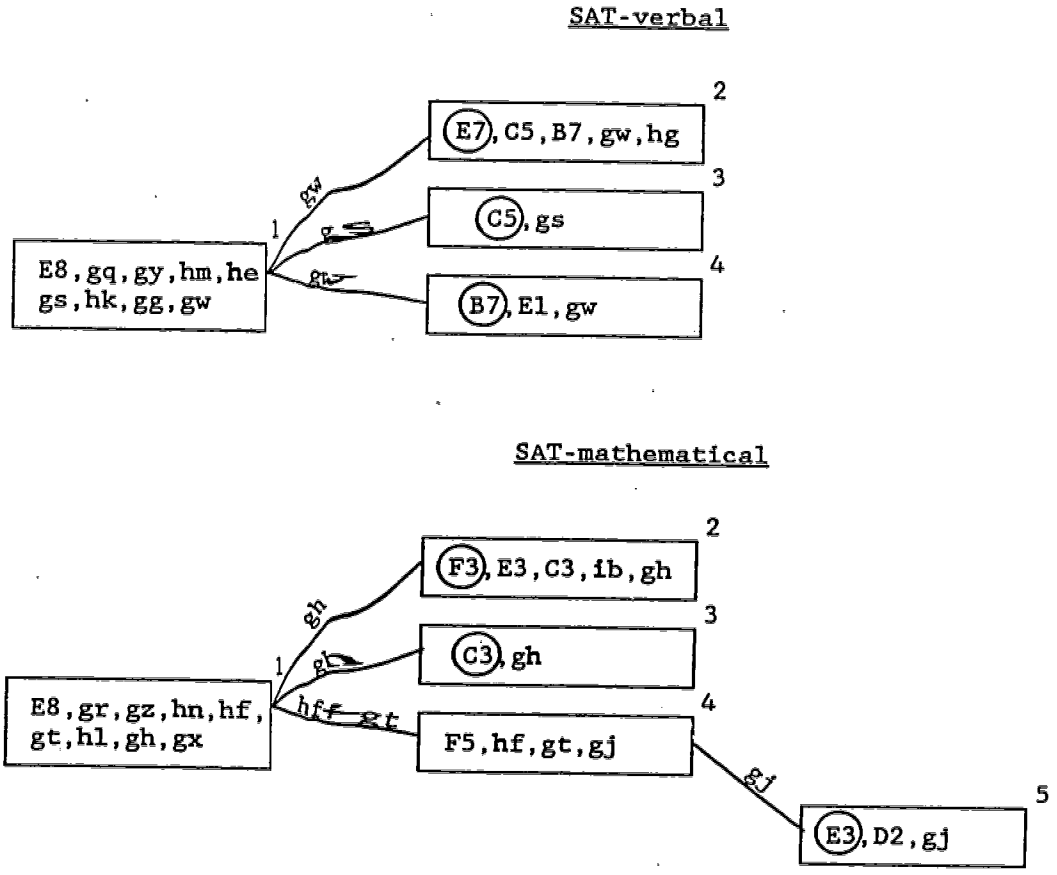
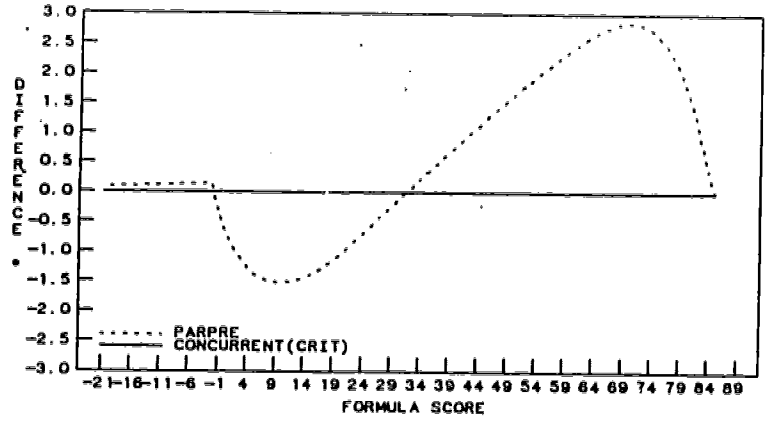


Figure 3: Portions of SAT-verbal and SAT-mathematical partial pre-calibration transformation plans containing specific editions under investigation in this study. Upper case letters and numbers designate operational editions; lower case letters designate equating sections. Parameter estimates for the partial pre-calibration equatings came from the editions that are circled. Numbers identify specific calibration runs.

Figure 4

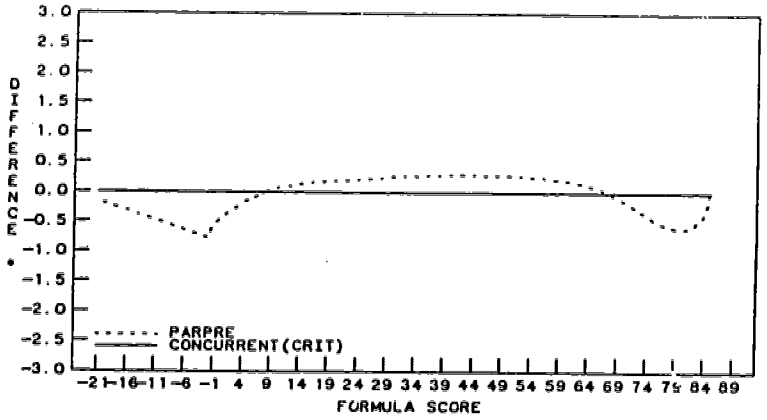
Plots of Raw-Score Equating Differences Derived from a Comparison of Partial Pre-calibration Equating Results to Concurrent Equating Results for SAT-verbal and SAT-mathematical Editions Being Studied[1]

SAT-verbal



DIFFERENCE = LINE - CRITERION LINE

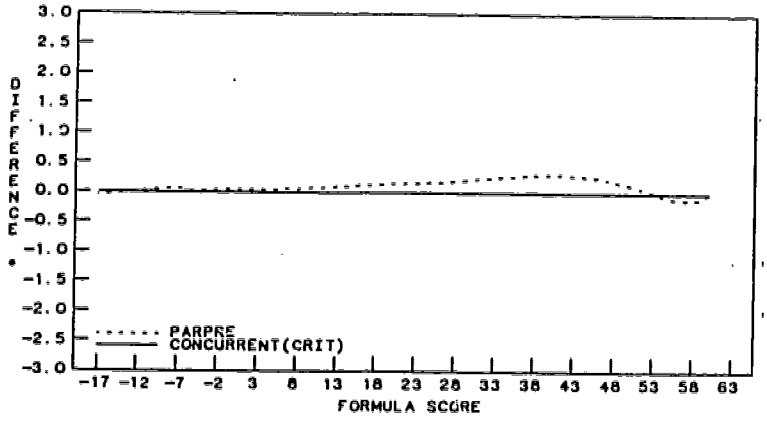
new edition: E7
old edition: C5



DIFFERENCE = LINE - CRITERION LINE

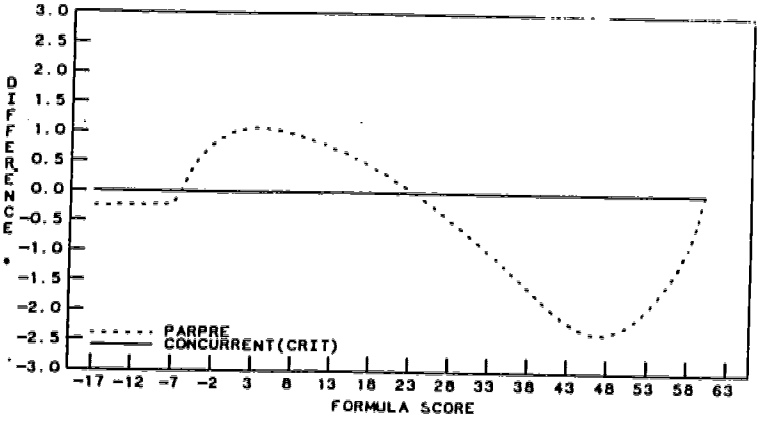
new edition: E7
old edition: B7

SAT-mathematical



DIFFERENCE = LINE - CRITERION LINE

new edition: F3
old edition: C3



DIFFERENCE = LINE - CRITERION LINE

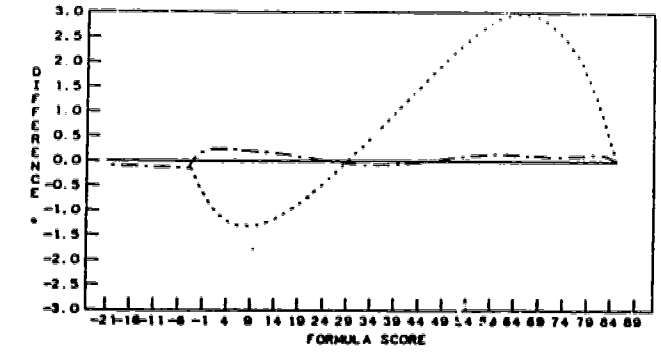
new edition: F3
old edition: E3

[1]Partial pre-calibration and concurrent equating results taken from Cook et al. (1985) study.

Figure 5

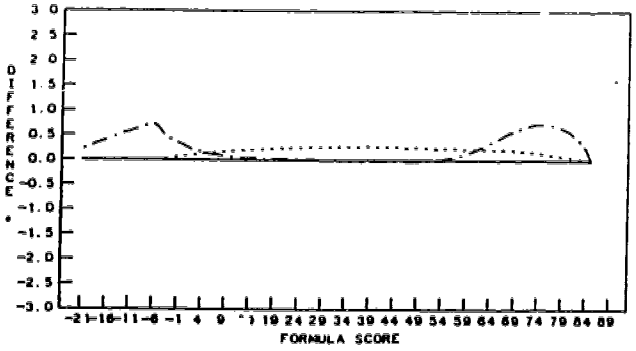
Plots of Raw-Score Equating Differences Derived from a Comparison of Special Equating Results to the Identity Transformation for SAT-verbal and SAT-mathematical Old Form Editions Being Studied[1]

SAT-verbal



* DIFFERENCE = LINE - CRITERION LINE

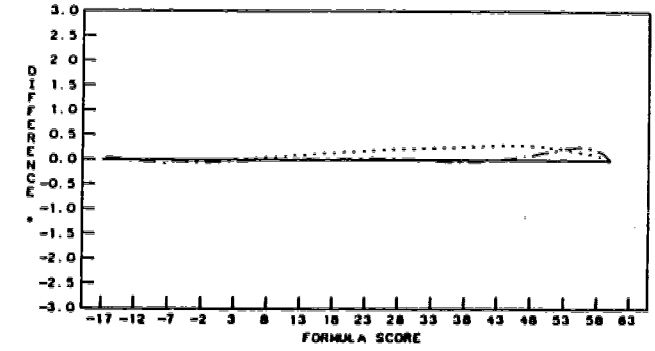
old edition: C5



* DIFFERENCE = LINE - CRITERION LINE

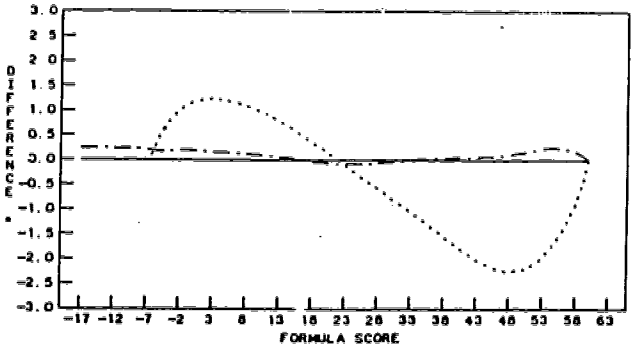
old edition: B7

SAT-mathematical



* DIFFERENCE = LINE - CRITERION LINE

old edition: C3



* DIFFERENCE = LINE - CRITERION LINE

old edition: E3

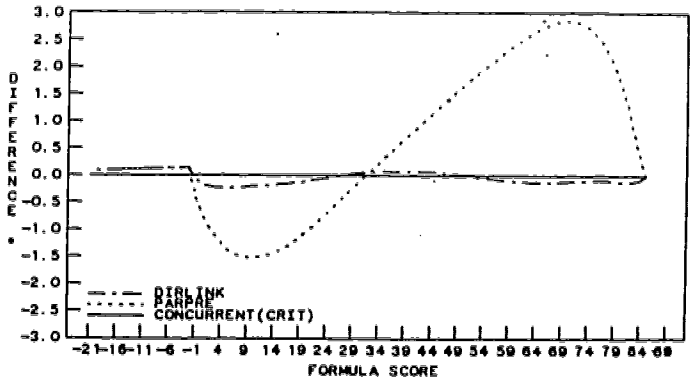
------- SPECIAL EQUATING 1
— – — SPECIAL EQUATING 2
——— CRITERION

[1] Special equatings are defined in Table 1 (see Equating 1 and Equating 2).

Figure 6

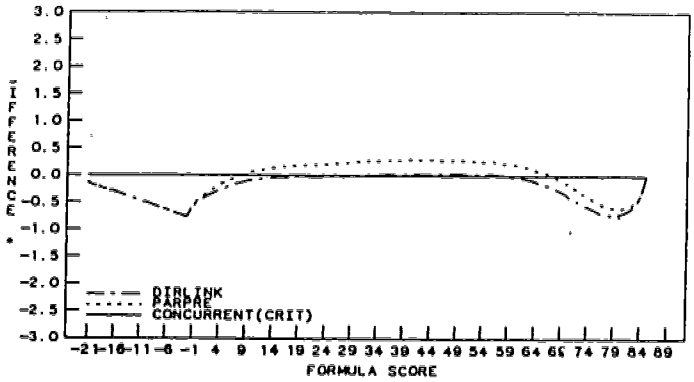
Plots of Raw-Score Equating Differences Derived from a Comparison of Partial Pre-calibration Equating Results and Direct Link Equating Results to Concurrent Equating Results for SAT-verbal and SAT-mathematical Editions Being Studied[1]

SAT-verbal



* DIFFERENCE = LINE - CRITERION LINE

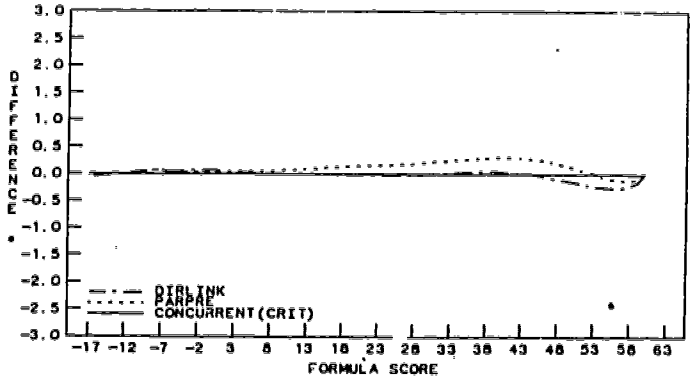
new edition: E7
old edition: C5



* DIFFERENCE = LINE - CRITERION LINE

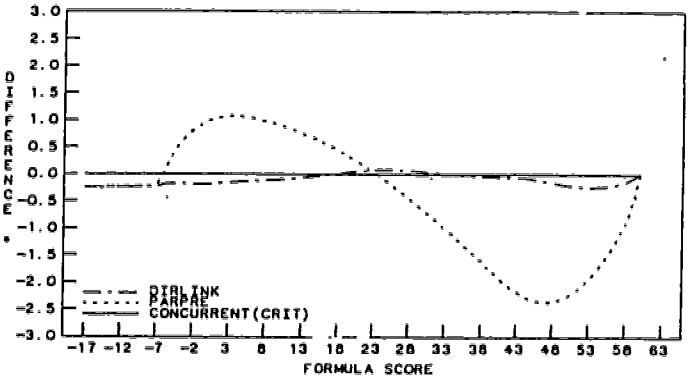
new edition: E7
old edition: B7

SAT-mathematical



* DIFFERENCE = LINE - CRITERION LINE

new edition: F3
old edition: C3



* DIFFERENCE = LINE - CRITERION LINE

new edition: F3
old edition: E3

[1] Partial pre-calibration and concurrent equating results taken from Cook et al. (1985) study.

Figure 7

Stem and Leaf Diagrams of Mean Absolute Differences (MAD) Between Item Response Functions for SAT-verbal Equating Sections

Equating Section gw Linking Runs 1 and 2		Equating Section gs Linking Runs 1 and 3		Equating Section gw Linking Runs 1 and 4	
.10		.10		.10	
.09		.09		.09	
.09		.09		.09	
.08		.08		.08	
.08	0	.08		.08	
.07		.07	6	.07	
.07		.07		.07	
.06		.06		.06	
.06		.06		.06	
.05		.05		.05	
.05		.05		.05	
.04		.04	5	.04	9
.04	1	.04		.04	
.03		.03	68	.03	
.03		.03	013	.03	0033
.02	688	.02	56	.02	
.02	01224	.02	000234	.02	233
.01	5566678999	.01	5667777889	.01	55677788899
.01	11122334444	.01	0001224	.01	0001122233
.00	5677789	.00	567778	.00	57777999999
.00	11	.00	14	.00	
40 items		40 items		40 items	

Figure 8

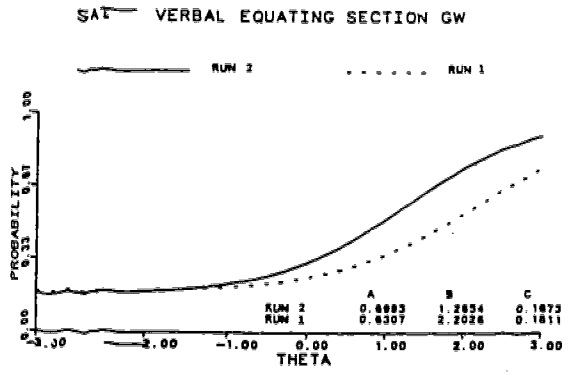
Stem and Leaf Diagrams of Mean Absolute Differences (MAD) Between Item Response Functions for SAT-mathematical Equating Sections

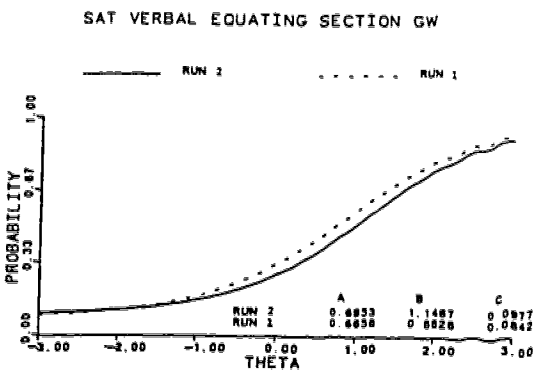
Equating Section gh Linking Runs 1 and 2		Equating Section gh Linking Runs 1 and 3		Equating Section hf + gt Linking Runs 1 and 4		Equating Section gj Linking Runs 4 and 5	
.10		.10	0	.10		.10	
.09		.09		.09		.09	
.09		.09		.09		.09	
.08		.08		.08		.08	
.08		.08		.08		.08	
.07		.07		.07		.07	
.07		.07		.07		.07	
.06		.06		.06		.06	
.06	1	.06		.06		.06	
.05		.05		.05		.05	
.05		.05		.05	23	.05	
.04		.04		.04		.04	
.04	2	.04		.04	3	.04	
.03		.03	57	.03	557	.03	
.03	02	.03	0	.03	1	.03	4
.02	6	.02	55899	.02	7	.02	589
.02	000	.02	123	.02	01111224	.02	000114
.01	55566677	.01	556678	.01	555566666888999	.01	577
.01	24	.01	0014	.01	000134	.01	0012
.00	67777	.00	79	.00	66677889	.00	5778
.00	23	.00	3	.00	2	.00	2334
25 items		25 items		50 items		25 items	

Figure 9

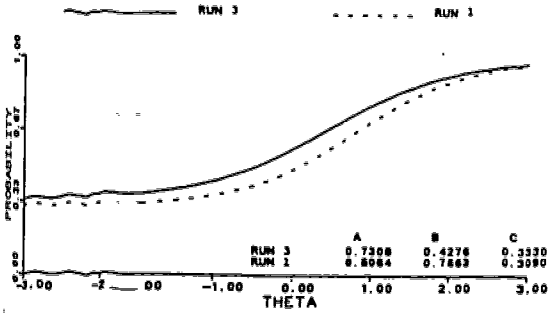
Plots of Item Response Functions for Items Removed from SAT-verbal Equating Sections

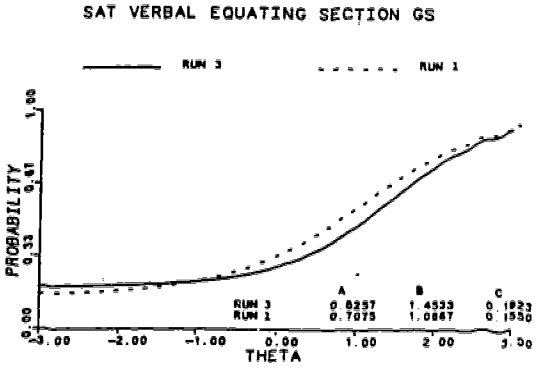
Equating Section gw - Linking Runs 1 and 2

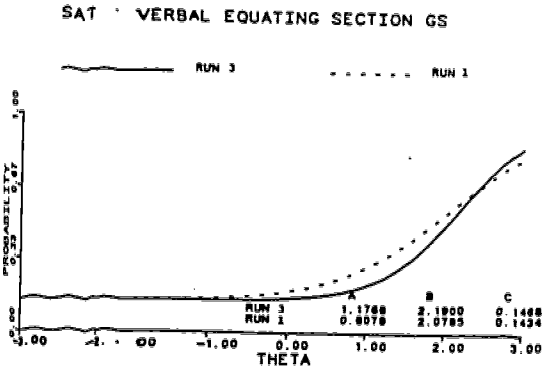


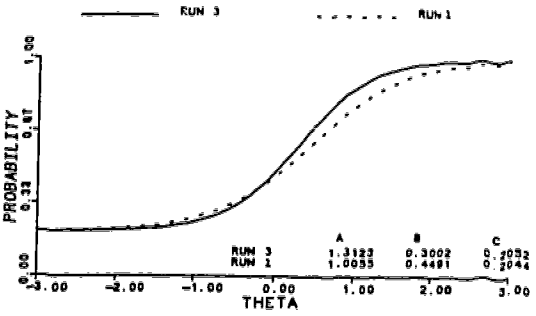


Equating Section gs - Linking Runs 1 and 2









Equating Section gw - Linking Runs 1 and 4

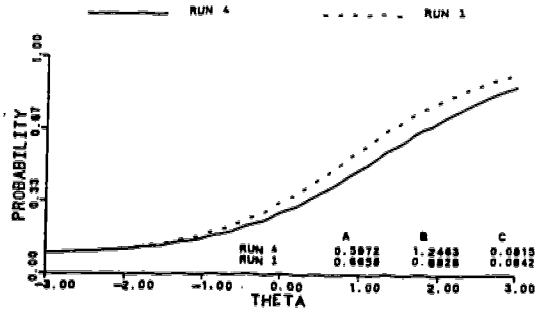
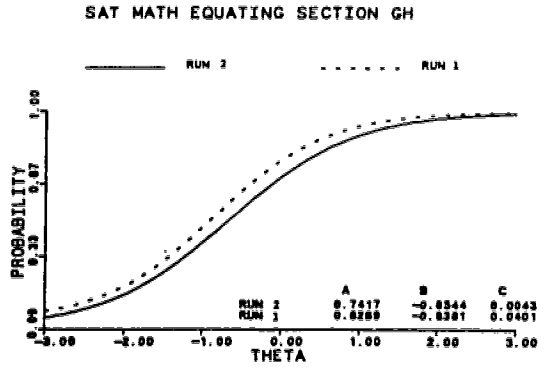
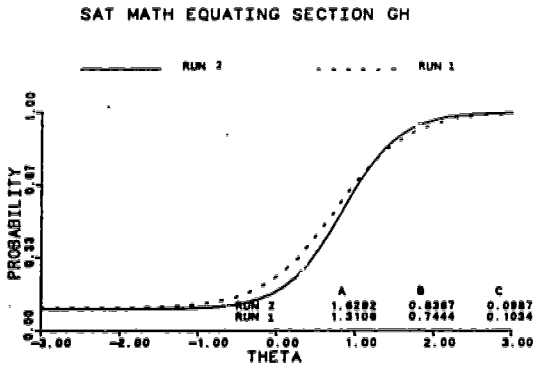


Figure 10

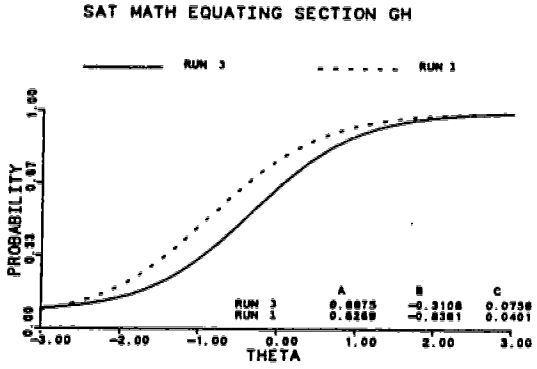
Plots of Item Response Functions for Items Removed from SAT-mathematical Equating Sections

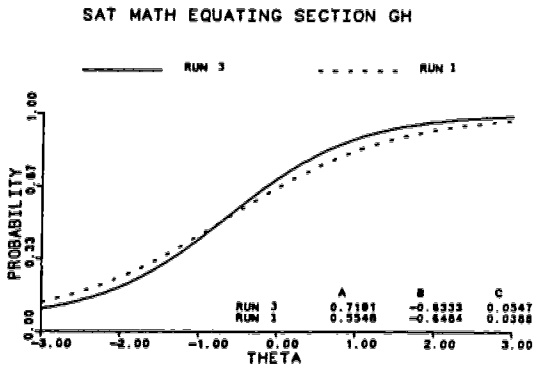
Equating Section gh - Linking Runs 1 and 2





Equating Section gh - Linking Runs 1 and 3





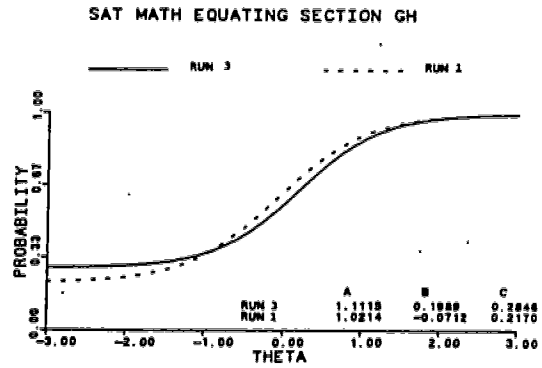
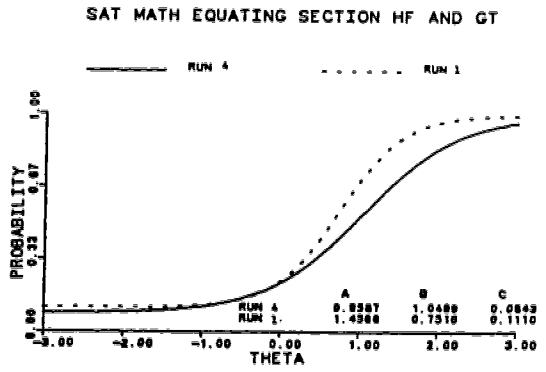
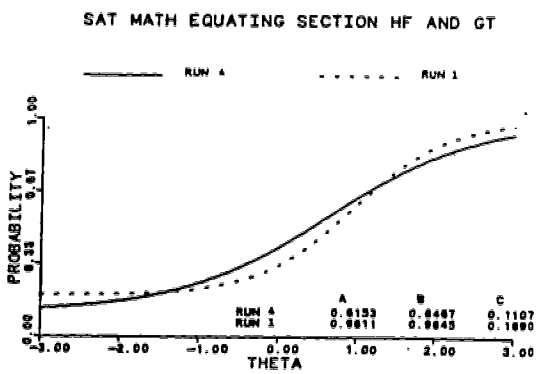


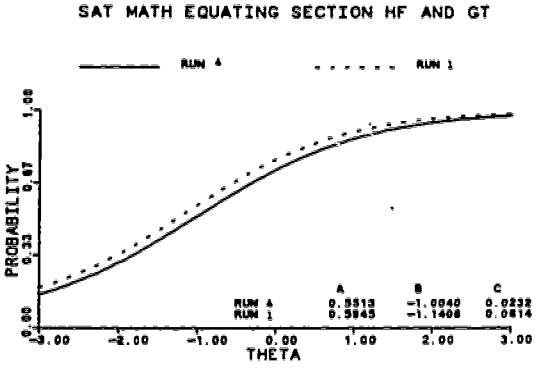
Figure 10 (continued)

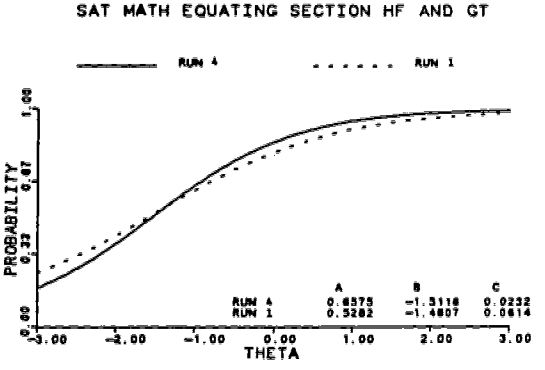
Plots of Item Response Functions for Items Removed from SAT-mathematical Equating Sections

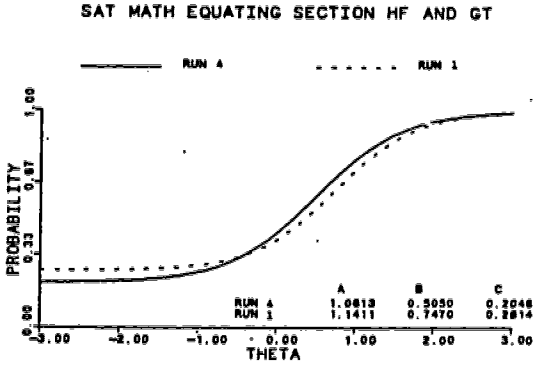
Equating Sections hf + gt - Linking Runs 1 and 4











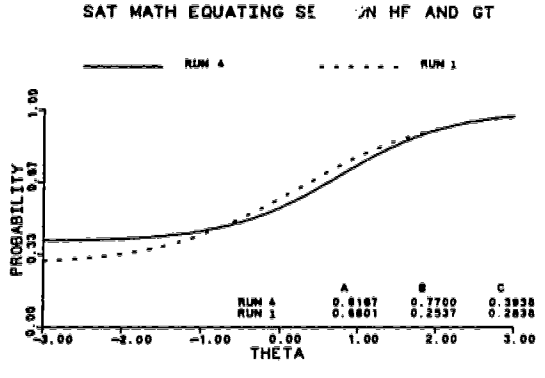
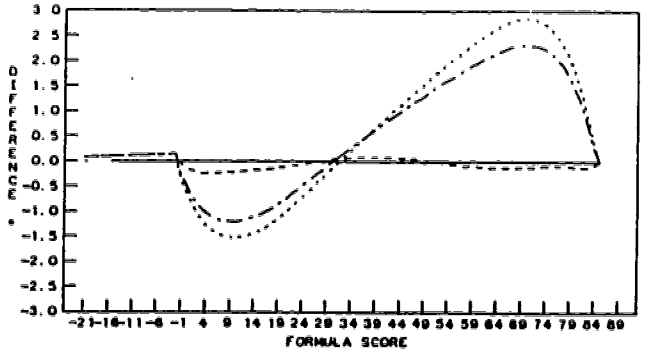


Figure 11

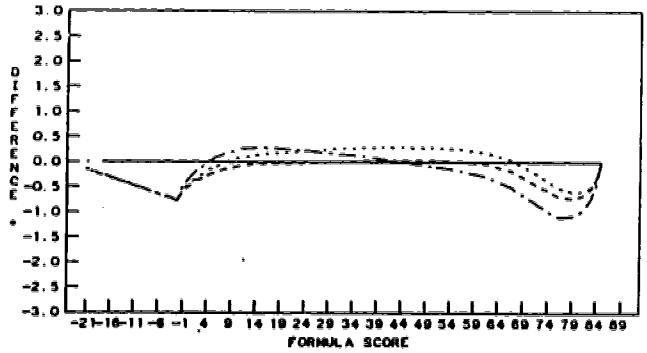
Plots of Raw-Score Equating Differences Derived from a Comparison of Previous Partial Pre-calibration, Current Partial Pre-calibration and Direct Link Equating Results to Concurrent Equating Results for SAT-verbal and SAT-mathematical Editions Being Studied[1]

SAT-verbal



* DIFFERENCE = LINE - CRITERION LINE

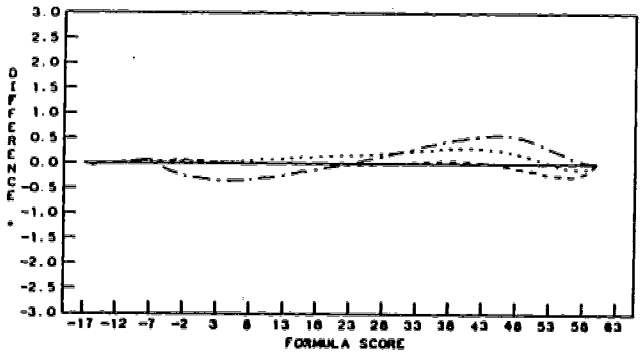
new edition: E7
old edition: C5



* DIFFERENCE = LINE - CRITERION LINE

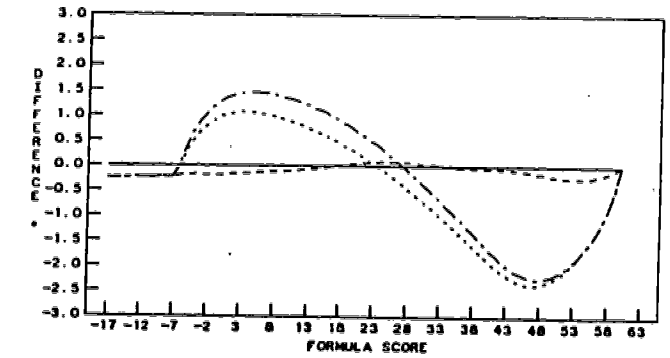
new edition: E7
old edition: B7

SAT-mathematical



* DIFFERENCE = LINE - CRITERION LINE

new edition: F3
old edition: C3



* DIFFERENCE = LINE - CRITERION LINE

new edition: F3
old edition: E3

---------- PREVIOUS PARTIAL PRE-CALIBRATION
— — — CURRENT PARTIAL PRE-CALIBRATION
— — — — DIRECT LINK
———— CONCURRENT CRITERION

[1]Previous partial pre-calibration results were taken from Cook et al. (1985) equatings. Current partial pre-calibration results involve same editions and linkings as previous results except that items exhibiting DIF have been removed from common item linking sections. Direct link equatings are described in Table 1.

Table 1

Summary of Special Equatings Performed to Study Linking Problems and Estimation Errors[1]

Old Edition	SAT-Verbal		SAT-Mathematical	
	C5	B7	C3	E3
Transformation	C5 (run 3)[2] placed on scale of run 2 using C5 items, i.e., C5 (run 3)*[3]	B7 (run 4) placed on scale of run 2 using B7 items, i.e., B7 (run 4)*	C3 (run 3) placed on scale of run 2 using C3 items, i.e., C3 (run 3)*	E3 (run 5) placed on scale of run 2 using E3 items, i.e., E3 (run 5)*
Parameter estimates for Equating 1	C5 (run 3)* to C5 (run 3)	B7 (run 4)* to B7 (run 4)	C3 (run 3)* to C3 (run 3)	E3 (run 5)* to E3 (run 5)
Parameter estimates for Equating 2	C5 (run 3)* to C5 (run 2)	B7 (run 4)* to B7 (run 2)	C3 (run 3)* to C3 (run 2)	E3 (run 5)* to E3 (run 2)
Parameter estimates for Direct Link Equating	E7 (run 2) to C5 (run 3)*	E7 (run 2) to B7 (run 4)*	F3 (run 2) to C3 (run 3)*	F3 (run 2) to E3 (run 5)*

[1] Parameter estimates for all editions used in these equatings have already been placed on the base scale (run 1 in Figure 3) as part of the Cook et al. (1985) study.

[2] The run, identified in Figure 3, from which the parameter estimates were taken is identified in parentheses, i.e., C5 parameter estimates from run 3 in Figure 3.

[3] The asterisk indicates that the parameter estimates have been transformed to the scale of a different calibration run identified in Figure 3.

Table 2

Linear Parameters Obtained from Direct Link
Item Parameter Transformations

Test	New Edition	Base Edition	Common Items	Linear Parameters Slope	Linear Parameters Intercept
SAT-verbal	B7 (run 4)	B7 (run 2)	B7	.9996	.0183
SAT-verbal	C5 (run 3)	C5 (run 2)	C5	1.0931	.0373
SAT-mathematical	E3 (run 5)	E3 (run 2)	E3	.9042	-.0278
SAT-mathematical	C3 (run 3)	C3 (run 2)	C3	1.0073	.0148

Table 3

Equating Section Summary Data for Adjacent SAT-verbal and SAT-mathematical Calibrations

SAT-verbal

Equating Section	Calibration Run[1]	Equating Section		Calibration Run	Equating Section	
		Mean	S.D.		Mean	S.D.
gw	1	16.53	7.80	2	16.87	8.07
gs	1	16.11	7.82	3	16.08	7.95
gw	1	16.53	7.80	4	14.07	7.90

SAT-mathematical

Equating Section	Calibration Run[1]	Equating Section		Calibration Run	Equating Section	
		Mean	S.D.		Mean	S.D.
gh	1	8.85	5.54	2	8.41	5.75
gh	1	8.85	5.54	3	8.40	5.51
hf+gt	1	19.68	----[2]	4	20.01	----[2]
gj	4	9.91	6.11	5	9.23	6.14

[1]Refers to specific calibration run identified in Figure 3.

[2]Could not be calculated from available data.

Appendix

Partial Pre-calibration Transformation Plan/Linkage System

An elaborate linkage system was devised in the Cook et al. (1985) study to allow placement of item parameter estimates on a common scale so that IRT equating resulting from a partial pre-calibration design could be investigated.

Figures A-1 and A-2 illustrate the design used for the verbal and mathematical sections, respectively. Each figure depicts the linkages necessary to place nine new editions, fifteen old editions, and associated equating tests on the base scale. It should be noted that upper case letter and number combinations indicate operational sections of the SAT, lower case letters indicate equating sections, boxes with solid lines enclose old test editions, and boxes with dotted lines indicate new test editions.

Lower case letter combinations, or occasionally, upper case letter and number combinations indicated above the arrows in Figure A-1 and A-2, denote common items that were used to place item parameter estimates from separate calibrations on a common scale via the characteristic curve transformation procedure (Stocking and Lord, 1983).

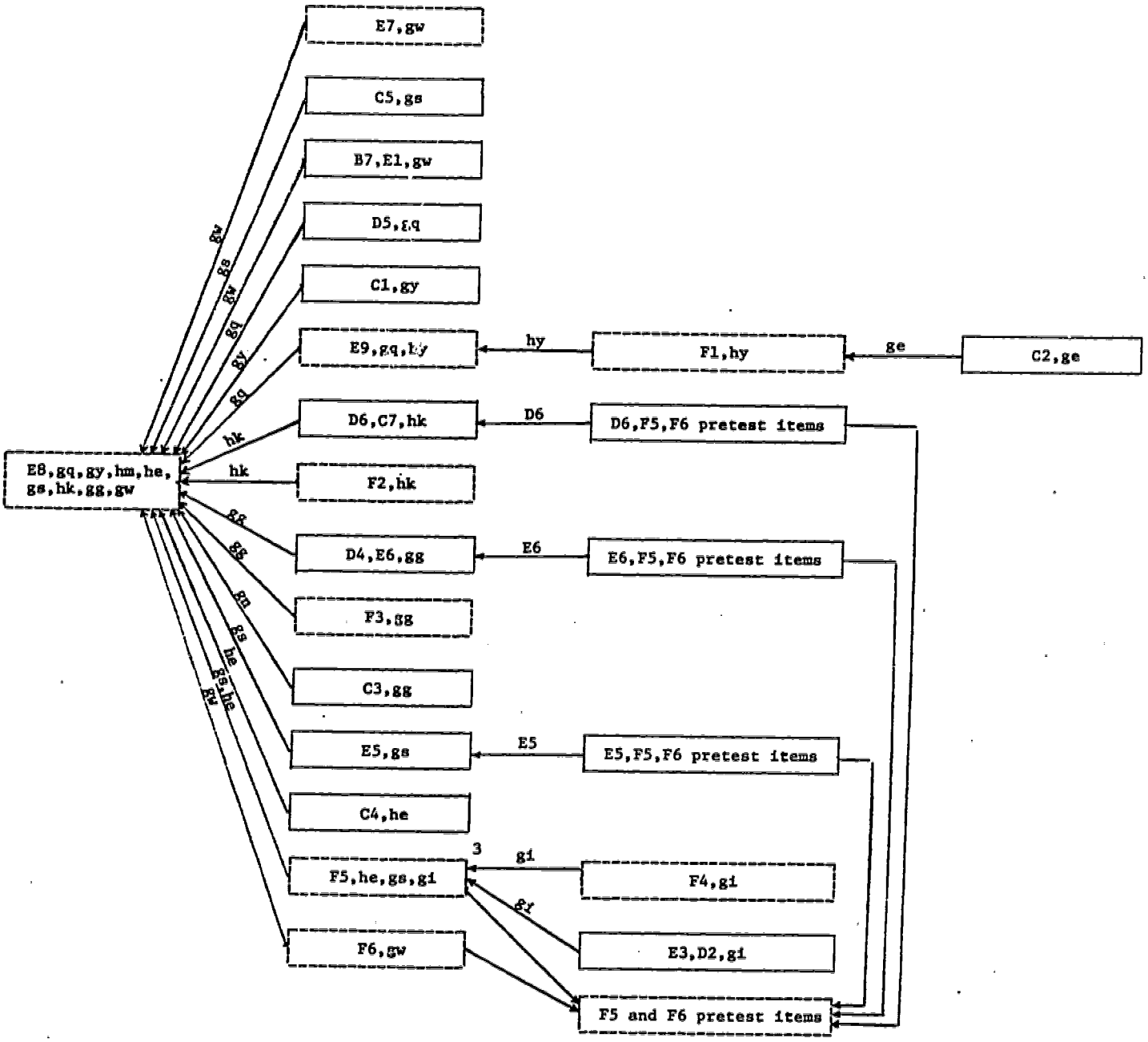
For SAT-verbal, the calibration run containing new edition E8 and eight equating sections, which were administered in November 1982, was used as the base item parameter scale to which all other sets of item parameter estimates were scaled. For example, item parameter estimates on a common scale exist for new verbal edition E7 and equating test gw from a previous concurrent calibration run (which is fully depicted in Figure 3 in the text). Item parameter estimates for gw also existed from the calibration run used as the base scale. The item parameter estimates for gw that were on scale with those for verbal edition E7 were scaled to those that are on the base E8 scale (using the characteristic curve transformation method). The resulting transformation was then applied to

the item parameter estimates for verbal edition E7 to convert them to the verbal base scale. The item parameter estimates for each new and old edition listed in Figure 2 were converted to the base scale in a similar manner to that just described for verbal edition E7. For some editions, more than one scaling was required to convert the item parameter estimates to the base scale. For example, item parameter estimates for edition F1 were converted to the same scale as those of edition E9 through equating test hy, which was then converted to the base scale through equating test gq. Item parameter estimates for new editions F5 and F6 were placed on scale in several different ways in the Cook et al. (1985) study. Those for verbal edition F5 were placed on scale using items from either equating section he or equating section gs or from the pooled he and gs equating sections. In addition, new editions F5 and F6 were used to study the effect of both full pre-calibration and pre-equating in the following way. Approximately 50% of the items contained in both editions were placed on the base scale as pretest items administered with test editions D6, E5, and E6. The remaining 50% of the items were placed on scale when they appeared in the final editions of F5 and F6, which were calibrated at their respective initial administrations. Item parameter estimates from these different calibrations were assembled (after they had been placed on the E8 scale) into pre-calibrated editions F5 and F6. The editions were then equated (simulating pre-equating) to their respective old editions.

The linking plan for SAT-mathematical (depicted in Figure A-2) is virtually identical to the plan previously described for SAT verbal. The only difference between Figures A-1 and A-2 are the lower case letters used to designate the equating sections.

Figure A.1

SAT-Verbal Transformation Plan[1,2]



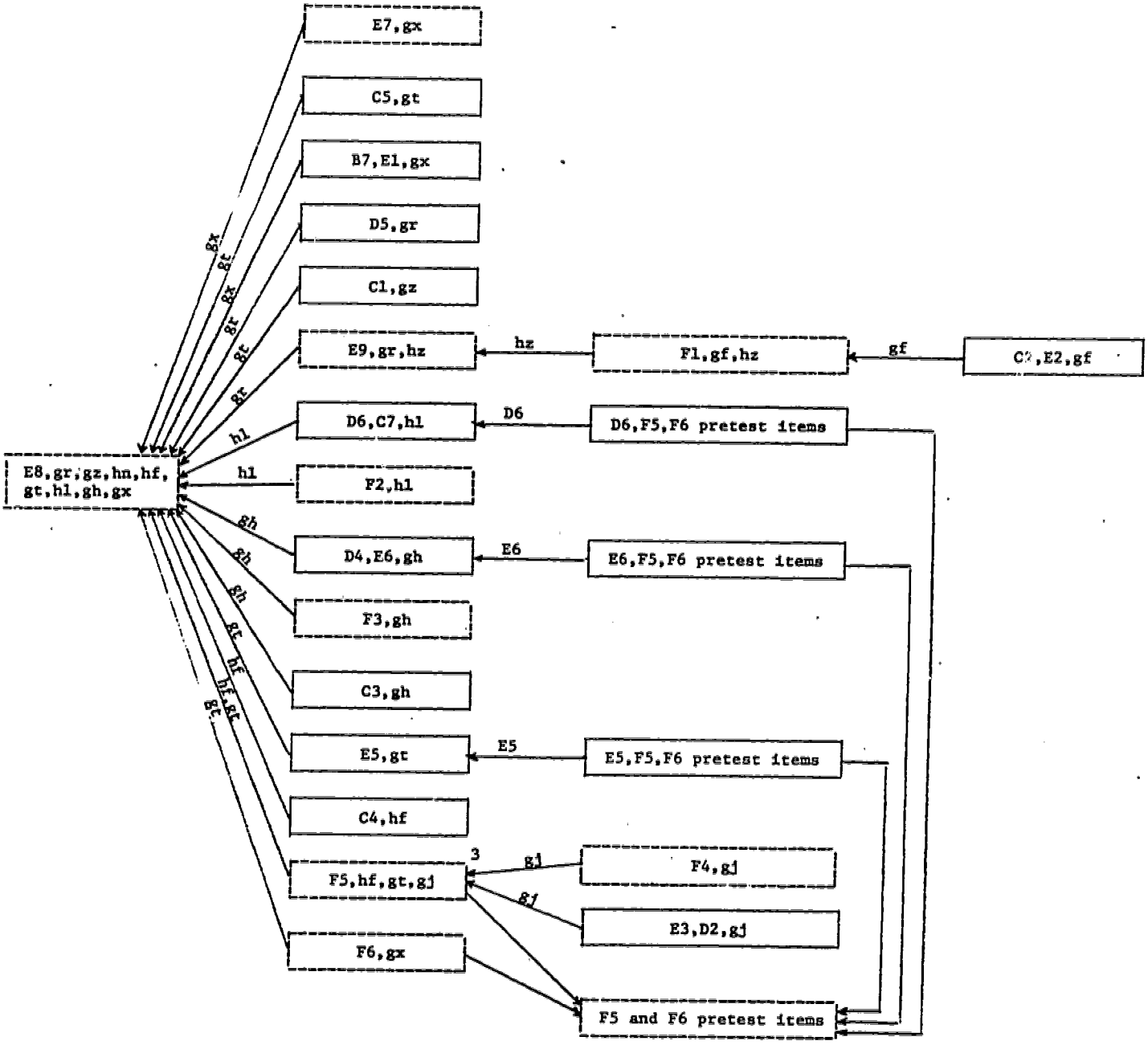
[1]Upper case letters and numbers designate operational editions; lower case letters designate equating sections.

[2]Dotted lines indicate new editions, solid lines indicate old editions.

[3]Edition F5 was placed on the E8 scale three ways: (1) using items contained in equating section he; (2) using items contained in equating section gs; and (3) using items contained in pooled equating sections he and gs.

Figure A.2

SAT-Mathematical Transformation Plan[1,2]



[1]Upper case letters and numbers designate operational editions; lower case letters designate equating sections.

[2]Dotted lines indicate new editions; solid lines indicate old editions.

[3]Edition F5 was placed on the E8 scale three ways: (1) using items contained in equating section hf; (2) using items contained in equating section gt; and (3) using items contained in pooled equating sections hf and gt.